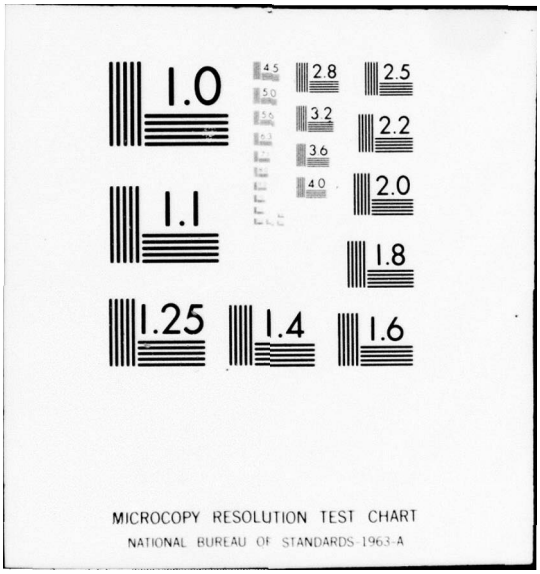
AD-A054 555 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
SOME OPTIMAL PROPERTIES AND INTERPRETATIONS OF PRINCIPAL COMPON--ETC(U)
MAR 78 R HUDLET, R A JOHNSON DAAG29-75-C-0024
UNCLASSIFIED MRC-TSR-1836 NL

1 OF 1
AD
A054555

END
DATE
FILMED
7-78
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

MRC Technical Summary Report #1836

SOME OPTIMAL PROPERTIES AND INTERPRETATIONS OF PRINCIPAL COMPONENTS

Raul Hudlet and Richard A. Johnson

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

March 1978

(Received December 21, 1977)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, DC 20550

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

SOME OPTIMAL PROPERTIES AND INTERPRETATIONS OF PRINCIPAL COMPONENTS

Raul Hudlet and Richard A. Johnson

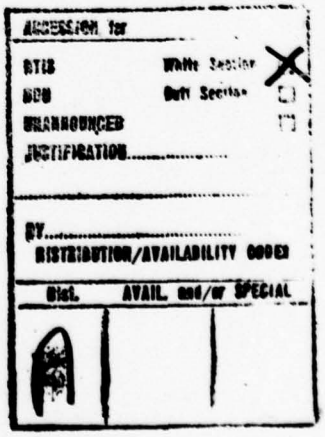
Technical Summary Report # 1836
March 1978

ABSTRACT

Optimal properties are derived and some new geometrical interpretations given for principal components. Typically, our main results concern the simultaneous minimization of eigenvalues of certain covariance matrices which measure the goodness of an approximation. Many popular criteria like total variance and generalized variance, which are increasing functions of the eigenvalues, are then minimized by the best approximator.

In other situations, the criterion may not be a monotone function of the eigenvalues. In Theorem 3.2, we derive a general optimal class based on the non-negative definite ordering of covariance matrices. Theorem 4.1 gives a result for the sequential selection of principal components. In the final section, we give a new geometrical interpretation of the sample principal components.

AMS (MOS) Subject Classification: 62H25
Key Words: Principal components, Statistical approximations
Work Unit Number 4 (Probability, Statistics, and Combinatorics)



Sponsored by the United States Army under Contract No. DAAG29-75-C-0024
and the National Science Foundation under Grant No. MCS77-09574.

SIGNIFICANCE AND EXPLANATION

When a large number of characteristics are measured on a large number of population units, the sheer volume of data can cause problems. It is natural, in such cases, to look for ways to reduce that data to a more manageable form.

One way of doing this is the Principal Component Method, wherein the original problem is replaced by an approximation of far lower dimension. The variables in the approximate problem are certain linear combinations, or principal components, of the variables in the original problem.

Of course, some choices of the exact linear combinations to be used (approximators) will yield more meaningful results than others. Ideally, we would like to retain as much information as possible, and we seek a set of principal components which is optimal from this point of view.

In this paper, we derive some new properties of optimal principal component approximators and obtain some further geometrical insights concerning this method. We also help clarify a potential weakness in this method by determining an even more general class of approximators, that contains those selected by the principal component, and exhibiting a situation where the approximator that would be selected by that method is not optimal.

SOME OPTIMAL PROPERTIES AND INTERPRETATIONS OF PRINCIPAL COMPONENTS

Raul Hudlet and Richard A. Johnson

1. Review of Previous Work

Let $\underset{\sim}{X} = (X_1,\ldots,X_p)'$ be a random vector with zero expectation and covariance matrix Σ and let $\lambda_1 \geq \ldots \geq \lambda_p$ denote the eigenvalues of Σ and $\underset{\sim}{P}_1,\ldots,\underset{\sim}{P}_p$ a corresponding set of orthonormal eigenvectors. Notice that $\underset{\sim}{P}_1,\ldots,\underset{\sim}{P}_p$ are uniquely determined (up to multiplication by ± 1) only if $\lambda_1 > \ldots > \lambda p$.

Even though Pearson had encountered principal components as early as 1901, the concept is generally attributed to Hotelling (1933) who was the first to introduce it in a probabilistic framework. Since then Girshick (1936), Anderson (1958), Rao (1964), Darroch (1965), Okamoto and Kanazawa (1968) among others have characterized principal components by different sets of optimality properties.

Girshick (1936) showed that if the components of $\underset{\sim}{X}$ have variance one, then $\underset{\sim}{P}_1'\underset{\sim}{X}$, a first principal component (P.C.), maximizes the sum of the squares of the correlation between $\underset{\sim}{\ell}'\underset{\sim}{X}$ and each variate X_i over all possible linear functions $\underset{\sim}{\ell}'\underset{\sim}{X}$.

Anderson (1958) established that among the class of linear functions $\underset{\sim}{\ell}'\underset{\sim}{X}$ with $\ell'\ell = 1$; a first P.C., $\underset{\sim}{P}_1'\underset{\sim}{X}$ has maximum variance; $\underset{\sim}{P}_2'\underset{\sim}{X}$, a second P.C., has maximum variance among the elements in the class uncorrelated with $\underset{\sim}{P}_1'\underset{\sim}{X}$ and so on.

Rao (1964) characterizes the first $k(\leq p)$ principal components as a linear form $\underset{\sim}{Y} = T\underset{\sim}{X}$, where T is a $k \times p$ matrix, which minimizes the trace or the Euclidean norm $(\|M\|^2 = \sum_{i,j} M_{ij}^2)$ of the covariance matrix of the residual of $\underset{\sim}{X}$ minus the best linear predictor based on $\underset{\sim}{Y}$.

Darroch (1965) was the first to characterize principal components within the class of all random variables with at most $k(\leq p)$ dimensions. In this formulation it is desired to approximate the p component vector $\underset{\sim}{X}$ by a linear form $A\underset{\sim}{Y}$ of a $k \times 1$ random vector $\underset{\sim}{Y}$ where A is a $p \times k$ matrix of constants. The error of approximation

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024 and the National Science Foundation under Grant No. MCS77-09574.

F, may be measured by the trace of the

$$\text{residual covariance matrix} = E(\underset{\sim}{X} - A\underset{\sim}{Y})(\underset{\sim}{X} - A\underset{\sim}{Y})' \ . \tag{1.1}$$

Darroch showed that F is minimized with respect to A and $\underset{\sim}{Y}$ when and only when

$$A\underset{\sim}{Y} = \underset{\sim}{P}_1(\underset{\sim}{P}_1'\underset{\sim}{X}) + \cdots + \underset{\sim}{P}_k(\underset{\sim}{P}_k'\underset{\sim}{X}) \tag{1.2}$$

which is a linear transformation of a set of k first principal components. Then, the minimum value is

$$F = \lambda_{k+1} + \cdots + \lambda_p$$

where, $\underset{\sim}{P}_1,\ldots,\underset{\sim}{P}_k$ is any set of k orthonormal eigenvectors of Σ corresponding, respectively, to the eigenvalues $\lambda_1,\ldots,\lambda_k$.

Okamoto and Kanazawa (1968) generalized Darroch's result, by allowing F to be any function of the eigenvalues of the residual covariance matrix, which is strictly increasing in each of its p arguments. Examples of such functions are the trace and the Euclidean norm. They showed that F is minimized when and only when Ay is as in (1.2), and then $F = F(\lambda_{k+1},\ldots,\lambda_p,0,\ldots,0)$. A nice review of these results appears in Okamoto (1969).

Our extensions concern the complete residual covariance matrix rather than just increasing functions of the eigenvalues. These latter results require invariance under orthogonal transformations and do not allow an investigator to single out special components of the observation vector. After first establishing some preliminary results in Section 2, we show in Section 3 that if the criterion for the fit is the non-negative definite partial ordering on the residual covariance matrix, then in general no overall optimal $A\underset{\sim}{Y}$ exists. However Theorem 3.2 establishes that an optimal class does exist. The implications of this extension are discussed in Section 3.3.

Principal components may also be introduced sequentially one at a time, by setting $k = 1$ above and demanding that at the jth stage, $A_j\underset{\sim}{Y}_j$, be uncorrelated with those selected at a previous stage and that the residual covariance matrix $E(\underset{\sim}{X} - A_j\underset{\sim}{Y}_j)(\underset{\sim}{X} - A_j\underset{\sim}{Y}_j)'$ have eigenvalues that are as small as possible. This is the approach taken in Section 4 leading to Theorem 4.1.

Section 5 illustrates how the results of Darroch (1965) and Okamoto and Kanazawa (1968) can be viewed as the population analoges to a generalization of Pearson's (1901) approach. Both Darroch (1965) and Okamoto and Kanazawa seem to have been unaware of this fact. We also obtain a result, Theorem 5.4, on the approximation of cross products matrices.

Lastly in Section 5.2 a seemingly new interpretation is given of the sample principal components in R^n.

2. A Preliminary Formulation

We are interested in approximating the vector $\underset{\sim}{X}$ by a random vector $A\underset{\sim}{Y}$ when goodness is measured by the covariance (1.1). Different choices of $A\underset{\sim}{Y}$ produce different residual covariance matrices and in order to compare them, we give the following definition.

A partial ordering in the class of all non-negative definite matrices A is defined by the relation $\geq$ where

$$A \geq B \quad \text{iff} \quad A - B \in \mathcal{A} . \tag{2.1}$$

The problem then becomes that of finding the $A\underset{\sim}{Y}$ which makes (1.1) as small as possible.

We begin by noticing that there is no loss in assuming $E\underset{\sim}{X} = \underset{\sim}{0}$. If this is not the case and $E\underset{\sim}{X} = \underset{\sim}{\mu}$ say, then similar to the regression situation where the best fitting polynomial is not forced to pass through the origin, we consider the residual to be defined by

$$E(\underset{\sim}{X} - \underset{\sim}{\eta} - A\underset{\sim}{Y})(\underset{\sim}{X} - \underset{\sim}{\eta} - A\underset{\sim}{Y})' \tag{2.2}$$

where $\underset{\sim}{\eta}$ is a $p \times 1$ unknown constant vector that one is allowed to vary when searching for a minimum.

If $E\underset{\sim}{Y} \neq \underset{\sim}{0}$, the term $AE\underset{\sim}{Y}$ may be absorbed into the $\underset{\sim}{\eta}$ by replacing $\underset{\sim}{X} - \underset{\sim}{\eta} - A\underset{\sim}{Y}$ by $\underset{\sim}{X} - (\underset{\sim}{\eta} + AE\underset{\sim}{Y}) - [A(\underset{\sim}{Y} - E\underset{\sim}{Y})]$. Since $\underset{\sim}{\eta}$ is arbitrary, $\underset{\sim}{\eta} + AE\underset{\sim}{Y}$ is also arbitrary and consequently we may restrict attention to $\underset{\sim}{Y}$ variables with zero expectation. Then

$$\begin{aligned} E(\underset{\sim}{X} - \underset{\sim}{\eta} - A\underset{\sim}{Y})(\underset{\sim}{X} - \underset{\sim}{\eta} - A\underset{\sim}{Y})' &= E(\underset{\sim}{X} - \underset{\sim}{\mu} - A\underset{\sim}{Y} - (\underset{\sim}{\eta} - \underset{\sim}{\mu}))(\underset{\sim}{X} - \underset{\sim}{\mu} - A\underset{\sim}{Y} - (\underset{\sim}{\eta} - \underset{\sim}{\mu}))' \\ &= E(\underset{\sim}{X} - \underset{\sim}{\mu} - A\underset{\sim}{Y})(\underset{\sim}{X} - \underset{\sim}{\mu} - A\underset{\sim}{Y})' + (\underset{\sim}{\eta} - \underset{\sim}{\mu})(\underset{\sim}{\eta} - \underset{\sim}{\mu})' \\ &\geq E(\underset{\sim}{X} - \underset{\sim}{\mu} - A\underset{\sim}{Y})(\underset{\sim}{X} - \underset{\sim}{\mu} - A\underset{\sim}{Y})' \end{aligned} \tag{2.3}$$

with strict inequality unless $\underset{\sim}{\eta} = \underset{\sim}{\mu}$.

Thus, in (2.3), we must take $\underset{\sim}{\eta} = \underset{\sim}{\mu}$. By assuming $\underset{\sim}{X}$ has already been corrected for its mean we get

$$E(\underset{\sim}{X} - A\underset{\sim}{Y})(\underset{\sim}{X} - A\underset{\sim}{Y})' \tag{2.4}$$

which is just the matrix in the original definition (1.1). Consequently, we may assume $E\underset{\sim}{X} = E\underset{\sim}{Y} = \underset{\sim}{0}$.

Also without loss of generality we may assume $E(\underset{\sim}{Y}\underset{\sim}{Y}') = I_k$, for even if rank of $E(\underset{\sim}{Y}\underset{\sim}{Y}') < k$ we may replace $A\underset{\sim}{Y}$ by $A^*\underset{\sim}{Y}^*$ where $\underset{\sim}{Y}^*$ is such that $E(\underset{\sim}{Y}^*\underset{\sim}{Y}^{*\prime}) = I_k$ and $A\underset{\sim}{Y} = A^*\underset{\sim}{Y}^*$ (a.s.).

Finally notice that if r = rank of Σ and $r \leq k$, the problem is trivial since we obtain a perfect fit with $A = [\underset{\sim}{P}_1,\ldots,\underset{\sim}{P}_r,\underset{\sim}{0},\ldots,\underset{\sim}{0}]$ and $\underset{\sim}{Y}' = (\underset{\sim}{P}_1'\underset{\sim}{X},\ldots,\underset{\sim}{P}_r'\underset{\sim}{X},\ldots,\underset{\sim}{P}_k'\underset{\sim}{X})$ so $A\underset{\sim}{Y} = \underset{\sim}{X}$ and

$$0 = E(\underset{\sim}{X} - A\underset{\sim}{Y})(X - A\underset{\sim}{Y})' \leq E(\underset{\sim}{X} - \bar{\bar{A}}\underset{\sim}{\bar{\bar{Y}}})(X - \bar{\bar{A}}\underset{\sim}{\bar{\bar{Y}}})' \tag{2.5}$$

with strict inequality for any choice of $\bar{\bar{A}}\underset{\sim}{\bar{\bar{Y}}}$ unless $AY = \bar{\bar{A}}\underset{\sim}{\bar{\bar{Y}}}$ (a.s.).

Thus in the rest of this paper, unless otherwise stated, we will assume that

$$E\underset{\sim}{X} = E\underset{\sim}{Y} = \underset{\sim}{0}, \quad E\underset{\sim}{Y}\underset{\sim}{Y}' = I_k, \quad k < r \ . \tag{2.6}$$

Lemma 2.1. Under assumptions (2.6) let

$$B = \operatorname{cov}(\underset{\sim}{X},\underset{\sim}{Y}) \ . \tag{2.7}$$

Then with strict inequality unless $B = A$

$$E(\underset{\sim}{X} - A\underset{\sim}{Y})(\underset{\sim}{X} - A\underset{\sim}{Y})' \geq \Sigma - BB' \geq 0 \ . \tag{2.8}$$

Proof.

$$E(\underset{\sim}{X} - A\underset{\sim}{Y})(\underset{\sim}{X} - A\underset{\sim}{Y})' = E(\underset{\sim}{X} - B\underset{\sim}{Y} + (B - A)\underset{\sim}{Y})(\underset{\sim}{X} - B\underset{\sim}{Y} + (B - A)\underset{\sim}{Y})' \geq E(\underset{\sim}{X} - B\underset{\sim}{Y})(\underset{\sim}{X} - B\underset{\sim}{Y})' = \Sigma - BB' \ . \quad \square$$

3. A Class of Best Approximators

The best current result, on the approximation of $\underset{\sim}{X}$ by $B\underset{\sim}{Y}$, is due to Okamoto and Kanazawa (1968). However, they restrict themselves to functions of eigenvalues of the residual matrix $E(\underset{\sim}{X} - B\underset{\sim}{Y})(\underset{\sim}{X} - B\underset{\sim}{Y})'$ which are increasing in each argument.

Theorem 3.1 [Okamoto and Kanazawa]. Let Σ have eigenvectors $\underset{\sim}{P}_1,\ldots,\underset{\sim}{P}_p$ corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Under assumptions (2.6), any strictly increasing function of the eigenvalues of

$$E(\underset{\sim}{X} - A\underset{\sim}{Y})(\underset{\sim}{X} - A\underset{\sim}{Y})'$$

is minimized with respect to A and $\underset{\sim}{Y}$ by the choice

$$A\underset{\sim}{Y} = \underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X} + \cdots + \underset{\sim}{P}_k\underset{\sim}{P}_k'\underset{\sim}{X} .$$

That is, the eigenvalues of the residual matrix are simultaneously minimized. □

Remark. One can conclude from Theorem 3.1 that the sum of residual variances (trace) is minimized, the generalized variance (det.) is minimized, or the sum of squares of all entries (sum of squared eigenvalues) is minimized.

We show below that dropping the invariance condition, implied by the restriction to functions of eigenvalues, leads to a whole class of optimal solutions derived from the partial ordering of non-negative definiteness.

3.1. Best Approximators

Lemma 2.1 tells us that, in trying to minimize (1.1) under assumptions (2.6), we can only improve the approximation if the matrix of A coefficients of $\underset{\sim}{Y}$ is chosen as the covariance between $\underset{\sim}{X}$ and $\underset{\sim}{Y}$. Now B must be found such that $\Sigma - BB'$ is as "small" as possible and then a $\underset{\sim}{Y}$ that produces such a B found. The next theorem will give us the structure of the non-negative definite covariance matrix $\Sigma - BB'$. It will be seen that there is no B which uniformly minimizes (2.8) but rather a collection of B's which are optimal in a sense.

Theorem 3.2. Let $\Sigma \geq 0$ be of order p and rank r and have eigenvectors $P = [\underset{\sim}{P}_1,\ldots,\underset{\sim}{P}_p]$ and eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Then under assumptions (2.6), set $\Lambda = \operatorname{diag}(\lambda_1,\ldots,\lambda_r)$ and

$$\Gamma_k = \{B^*\underset{\sim}{Y}^* \mid B^*\underset{\sim}{Y}^* = P\begin{bmatrix} \Lambda^{1/2}R\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}R'\Lambda^{-1/2} & 0 \\ 0 & 0 \end{bmatrix}P'\underset{\sim}{X};\ R \quad r \times r \quad \text{orthogonal}\}$$

so that by changing $R\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}R'$ the different elements of Γ_k are obtained. Then

(1) [Completeness of Γ_k]. Given any random vector $A\underset{\sim}{Y}$, there exists an element $B^*\underset{\sim}{Y}^*$ in the class such that

$$E(\underset{\sim}{X} - B^*\underset{\sim}{Y}^*)(\underset{\sim}{X} - B^*\underset{\sim}{Y}^*)' \leq E(\underset{\sim}{X} - A\underset{\sim}{Y})(\underset{\sim}{X} - A\underset{\sim}{Y})'$$

with strict inequality unless (a.s.) $A\underset{\sim}{Y} = B^*\underset{\sim}{Y}^*$.

(2) [Minimality of Γ_k]. Let $B\underset{\sim}{Y}$, in Γ_k, be given. There is no $\bar{B}\underset{\sim}{\bar{Y}}$ in Γ_k (with $\bar{B}\underset{\sim}{\bar{Y}}$ not a.s. equal to B^*Y^*) such that

$$E(\underset{\sim}{X} - \bar{B}\underset{\sim}{\bar{Y}})(X - \bar{B}\underset{\sim}{\bar{Y}})' \leq E(\underset{\sim}{X} - B^*\underset{\sim}{Y}^*)(\underset{\sim}{X} - B^*Y^*)' \ . \qquad \square$$

3.2. Proof of Theorem 3.2

We now proceed to develop a proof of Theorem 3.2 through a series of results.

Theorem 3.3. Let Σ of order p and rank r be n.n.d. and B be $p \times k$ such that $\Sigma - BB' \geq 0$. Then there exists an orthogonal matrix R of order r such that

$$BB' = P\begin{bmatrix} \Lambda^{1/2}R\begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}R'\Lambda^{1/2} & 0 \\ 0 & 0 \end{bmatrix}P'$$

where $D = \mathrm{diag}(d_1,\ldots,d_k)$; $1 \geq d_1 \geq \cdots \geq d_k \geq 0$; $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_r)$ and P is orthogonal of order p and its ith column $\underset{\sim}{P}_i$ is an eigenvector of Σ, corresponding to the ith largest eigenvalue of Σ, namely $\lambda_i (i = 1,\ldots,p)$.

Proof. By the spectral decomposition theorem, an orthogonal matrix P exists such that

$$\Sigma = P\begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix}P'$$

where Λ is a $r \times r$ diagonal matrix with diagonal entries $\lambda_1 \geq \cdots \geq \lambda_r \geq 0$. Then

$$\Sigma - BB' = P\begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix}P' - BB' \geq 0 \iff \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} - P'BB'P \geq 0 \ . \tag{3.1}$$

From (3.1) it is clear that the lower right hand corner of $P'BB'P$ must be zero, that is

$$P'B = \begin{bmatrix} C \\ 0 \end{bmatrix} \tag{3.2}$$

for some $r \times k$ matrix C. Equation (3.2) then takes the form

$$\begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} CC' & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \Lambda - CC' & 0 \\ 0 & 0 \end{bmatrix} \tag{3.3}$$

and we may thus restrict attention to the upper left corners.

$$\Lambda - CC' \geq 0 \iff I_r - \Lambda^{-1/2} CC' \Lambda^{-1/2} \geq 0 \ . \tag{3.4}$$

Again by the spectral theorem an orthogonal matrix R exists such that

$$R' \Lambda^{-1/2} CC' \Lambda^{-1/2} R = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \tag{3.5}$$

where $D = \operatorname{diag}(d_1, \ldots, d_k)$; $d_1 \geq \cdots \geq d_k$. Substituting in (3.4)

$$I_r - \Lambda^{-1/2} CC' \Lambda^{-1/2} \geq 0 \iff I_r - \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \geq 0 \ . \tag{3.6}$$

From (3.2)

$$B = P \begin{bmatrix} C \\ 0 \end{bmatrix}$$

so

$$BB' = P \begin{bmatrix} CC' & 0 \\ 0 & 0 \end{bmatrix} P' \ . \tag{3.7}$$

Equation (3.5) gives

$$CC' = \Lambda^{1/2} R \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} R' \Lambda^{1/2} \tag{3.8}$$

which in (3.7) yields

$$BB' = P \begin{bmatrix} \Lambda^{1/2} R \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} R' \Lambda^{1/2} & 0 \\ 0 & 0 \end{bmatrix} P' \ .$$

Finally, (3.6) shows $1 \geq d_1 \geq \cdots \geq d_k \geq 0$. □

Theorem 3.4 shows that the choice $D = I_k$ is best.

Theorem 3.4. Under the conditions of Theorem 3.3, let

$$B^*B^{*\prime} = P\begin{bmatrix} \Lambda^{1/2}R\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}R'\Lambda^{1/2} & 0 \\ 0 & 0 \end{bmatrix}P' . \tag{3.9}$$

Then

(i) $$0 \leq \Sigma - B^*B^{*\prime} \leq \Sigma - BB' \tag{3.10}$$

with strict inequality unless $B^*B^{*\prime} = BB'$.

(ii) (Admissibility of B^*). If $\bar{B}$ is an arbitrary $p \times k$ matrix such that

$$0 \leq \Sigma - \bar{B}\bar{B}' \leq \Sigma - B^*B^{*\prime} \tag{3.11}$$

then

$$\bar{B}\bar{B}' = B^*B^{*\prime} . \tag{3.12}$$

Proof. $(\Sigma - BB') - (\Sigma - B^*B^{*\prime}) = B^*B^{*\prime} - BB'$.

By Theorem 3.3 and (3.9), this can be written as

$$P\begin{bmatrix} \Lambda^{1/2}R\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}R'\Lambda^{1/2} & 0 \\ 0 & 0 \end{bmatrix}P' - P\begin{bmatrix} \Lambda^{1/2}R\begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}R'\Lambda^{1/2} & 0 \\ 0 & 0 \end{bmatrix}P'$$

$$= P\begin{bmatrix} \Lambda^{1/2}R\begin{bmatrix} I_k - D & 0 \\ 0 & 0 \end{bmatrix}R'\Lambda^{1/2} & 0 \\ 0 & 0 \end{bmatrix}P' \geq 0 \tag{3.13}$$

since, by (3.6), $1 \geq d_1 \geq \cdots \geq d_k \geq 0$. Notice the inequality is strict unless $D = I_k$ in which case $BB' = B^*B^{*\prime}$.

Suppose $\bar{B}$ exists such that (3.11) holds. Then, proceeding as in Theorem 3.3, $\bar{R}$ orthogonal and $\bar{D}$ diagonal exist (the notation being clear) such that

$$\bar{B}\bar{B}' = P\begin{bmatrix} \Lambda^{1/2}\bar{R}\begin{bmatrix} \bar{D} & 0 \\ 0 & 0 \end{bmatrix}\bar{R}'\Lambda^{1/2} & 0 \\ 0 & 0 \end{bmatrix}P' . \tag{3.14}$$

From (i) of the present theorem

$$\bar{B}^*\bar{B}^{*\prime} = P\begin{bmatrix} \Lambda^{1/2}\bar{R}\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}\bar{R}'\Lambda^{1/2} & 0 \\ 0 & 0 \end{bmatrix}P' \tag{3.15}$$

is such that

$$0 \leq \Sigma - \bar{B}^*\bar{B}^{*\prime} \leq \Sigma - \bar{B}\bar{B}' \leq \Sigma - B^*B^{*\prime} . \tag{3.16}$$

Equations (3.9) and (3.15) show that (3.16) is equivalent to

$$\bar{R}\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}\bar{R}' \geq R\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}R' . \tag{3.17}$$

It is easy to see that (3.17) can hold only if the equality holds, for it is clear that the first column of R, say r_1, is an eigenvector of the right hand side of (3.17) corresponding to the eigenvalue 1. Since the eigenvalues are 1's and 0's, by the Courant-Fischer minimax theorem,

$$r_1'r_1 = 1 \leq r_1'\bar{R}\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}\bar{R}'r_1 \leq 1 .$$

For equality, r_1 must be an eigenvector of the left hand side of (3.17) corresponding to the eigenvalue 1. Similarly it can be shown that the other columns of R are eigenvectors of both sides corresponding to the same eigenvalues and thus the two sides have the same eigenvalues and a common set of orthonormal eigenvectors which implies

$$\bar{R}\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}\bar{R}' = R\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}R' . \tag{3.18}$$

(Notice however this does not imply $R = \bar{R}$). □

The next lemma gives us the general form of B^*.

Lemma 3.5. If $B^*B^{*\prime}$ is as in (3.9) and B^* is $p \times k$ then B^* is of the form

$$B^* = P\begin{bmatrix} \Lambda^{1/2}R\begin{bmatrix} Q \\ 0 \end{bmatrix} \\ 0 \end{bmatrix} \tag{3.19}$$

where Q is an arbitrary orthogonal matrix of order k.

Proof. We can rewrite (3.9) as

$$B^*B^{*\prime} = P\begin{bmatrix} \Lambda^{1/2}R\begin{bmatrix} I_k \\ 0 \end{bmatrix} \\ 0 \end{bmatrix}[[I_k \quad 0]R'\Lambda^{1/2}, 0]P' \ . \tag{3.20}$$

Vinograd's theorem then implies that

$$B^* = P\begin{bmatrix} \Lambda^{1/2}R\begin{bmatrix} I_k \\ 0 \end{bmatrix} \\ 0 \end{bmatrix}Q = P\begin{bmatrix} \Lambda^{1/2}R\begin{bmatrix} Q \\ 0 \end{bmatrix} \\ 0 \end{bmatrix} \tag{3.21}$$

where Q is $k \times k$ orthogonal. □

It remains to find $\underset{\sim}{Y}^*$ which satisfies the conditions in (2.6) and such that $E(\underset{\sim}{X}\underset{\sim}{Y}^{*\prime}) = B^*$. Although the argument that follows is very similar to that of Darroch (1965), it is given for completeness.

Lemma 3.6. Let

$$H = P\begin{bmatrix} \Lambda^{-1/2}R\begin{bmatrix} Q \\ 0 \end{bmatrix} \\ 0 \end{bmatrix} \tag{3.22}$$

and B^* be as in (3.21). Define $\underset{\sim}{Y}^* = H'\underset{\sim}{X}$. Then Y^* is (a.s.) the only $k \times 1$ random vector satisfying

$$\begin{aligned} E(Y^*) &= 0 \\ E(\underset{\sim}{Y}^*\underset{\sim}{Y}^{*\prime}) &= I_k \\ E(\underset{\sim}{X}\underset{\sim}{Y}^{*\prime}) &= B^* \ . \end{aligned} \tag{3.23}$$

Proof. Since

$$\Sigma H = P\begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix}P' \cdot P\begin{bmatrix} \Lambda^{-1/2}R\begin{bmatrix} Q \\ 0 \end{bmatrix} \\ 0 \end{bmatrix} = P\begin{bmatrix} \Lambda^{1/2}R\begin{bmatrix} Q \\ 0 \end{bmatrix} \\ 0 \end{bmatrix} = B^*$$

and

$$H'B^* = [(Q' \vdots 0)R'\Lambda^{-1/2},\ 0]P'P\begin{bmatrix}\Lambda^{1/2}R\begin{bmatrix}Q\\0\end{bmatrix}\\0\end{bmatrix} = I_k \tag{3.24}$$

the conditions on the moments follow.

If $\underset{\sim}{Y}$ also satisfies (3.23), then

$$E(\underset{\sim}{Y} - H'\underset{\sim}{X})(\underset{\sim}{Y} - H'\underset{\sim}{X})' = I_k - E(\underset{\sim}{Y}\underset{\sim}{X}'H) - E(H'\underset{\sim}{X}\underset{\sim}{Y}') + I_k = 0 \tag{3.25}$$

so $\underset{\sim}{Y} = H'\underset{\sim}{X} = \underset{\sim}{Y}^*$ (a.s.).

Remark. It must be noticed that even though (3.21) shows that B^* is arbitrary up to a choice of Q so $\underset{\sim}{Y}^*$ in Lemma 3.6 may change, the product $B^*\underset{\sim}{Y}^*$ is invariant.

$$\begin{aligned} B^*\underset{\sim}{Y}^* &= P\begin{bmatrix}\Lambda^{1/2}R\begin{bmatrix}Q\\0\end{bmatrix}\\0\end{bmatrix}[(Q' \vdots 0)R'\Lambda^{-1/2}\ 0]P'\underset{\sim}{X} \\ &= P\begin{bmatrix}\Lambda^{1/2}R\begin{bmatrix}I_k & 0\\0 & 0\end{bmatrix}R'\Lambda^{-1/2} & 0\\ 0 & 0\end{bmatrix}P'\underset{\sim}{X} \end{aligned} \tag{3.26}$$

which does not depend on Q. We are now ready to prove our main result.

Proof (Theorem 3.2). (1) Let $A\underset{\sim}{Y}$ be given. With $B = \mathrm{cov}(\underset{\sim}{X},\underset{\sim}{Y})$, Lemma 2.1 establishes that

$$E(\underset{\sim}{X} - B\underset{\sim}{Y})(\underset{\sim}{X} - B\underset{\sim}{Y})' = \Sigma - BB' \leq E(\underset{\sim}{X} - A\underset{\sim}{Y})(\underset{\sim}{X} - A\underset{\sim}{Y})' \tag{3.27}$$

with strict inequality unless $B = A$.

Since $\Sigma - BB' \geq 0$ from (2.8), Theorem 3.4 gives B^* such that

$$0 \leq \Sigma - B^*B^{*\prime} \leq \Sigma - BB' \tag{3.28}$$

with strict inequality unless $B^*B^{*\prime} = BB'$. Next, let $\underset{\sim}{Y}^*$ be as in Lemma 3.6 so

$$E(\underset{\sim}{X} - B^*\underset{\sim}{Y}^*)(\underset{\sim}{X} - B^*\underset{\sim}{Y}^*)' = \Sigma - B^*B^{*\prime} \leq E(X - AY)(X - AY)' \tag{3.29}$$

with strict inequality unless $B^*B^{*\prime} = BB'$ and $B = A$. Suppose that equality holds. Then $B^*B^{*\prime} = BB'$ and by Lemma 3.5 applied to BB', there exists an orthogonal Q, of order k, such that

$$B = P\begin{bmatrix} \Lambda^{1/2}R\begin{bmatrix} Q \\ 0 \end{bmatrix} \\ 0 \end{bmatrix}. \tag{3.30}$$

Lemma 3.6 then implies that (a.s.)

$$\underset{\sim}{Y} = ((Q' \vdots 0)R'\Lambda^{-1/2}, 0)P'\underset{\sim}{X} \tag{3.31}$$

and by the remark after Lemma 3.6, $B^*\underset{\sim}{Y}^* = B\underset{\sim}{Y}$ (a.s.).

For the proof of (2), we construct $\bar{B}^*\bar{\underset{\sim}{Y}}^*$, just as B^*Y^* was constructed in part (1), such that

$$E(\underset{\sim}{X} - \bar{B}^*\bar{\underset{\sim}{Y}}^*)(\underset{\sim}{X} - \bar{B}^*\bar{\underset{\sim}{Y}}^*)' = \Sigma - \bar{B}^*\bar{B}^{*\prime} \leq E(\underset{\sim}{X} - B^*\underset{\sim}{Y}^*)(\underset{\sim}{X} - B^*\underset{\sim}{Y}^*)' \tag{3.32}$$

That is, $\Sigma - \bar{B}^*\bar{B}^{*\prime} \leq \Sigma - B^*B^{*\prime}$ but then part (ii) of Theorem 3.4 implies that $\bar{B}^*\bar{B}^{*\prime} = B^*B^{*\prime}$. As in part (i), it then follows that $\bar{B}^*\bar{\underset{\sim}{Y}}^* = B^*\underset{\sim}{Y}^*$ (a.s.).

3.3. Some Implications of Theorem 3.2

When predicting $\underset{\sim}{X}$ by $B\underset{\sim}{Y}$ with $\underset{\sim}{Y}$ $k \times 1$, if loss is measured by a function of the eigenvalues of the residual covariance matrix, then Okamoto and Kanazawa (1968) give conditions under which principal components are optimal. Other objectives, which are not expressible as functions of the eigenvalues, lead to the selection of other members of the class Γ_k.

To illustrate this point, we consider the case $k = 1$. Suppose that, for a given vector $\underset{\sim}{a}$, it is desired to find BY such that

$$\underset{\sim}{a}'E(\underset{\sim}{X} - BY)(\underset{\sim}{X} - BY)'\underset{\sim}{a} \tag{3.33}$$

is a minimum. This monotone, non-decreasing loss function over the class of non-negative definite matrices is not a function of the eigenvalues. Here

$$\underset{\sim}{a}'E(\underset{\sim}{X} - BY)(\underset{\sim}{X} - BY)'\underset{\sim}{a} = \underset{\sim}{a}'(\Sigma - BB')\underset{\sim}{a} =$$

$$\mathrm{Var}(\underset{\sim}{a}'\underset{\sim}{X}) - \mathrm{Cov}^2(\underset{\sim}{a}'\underset{\sim}{X}, \underset{\sim}{Y}) \geq \mathrm{Var}(\underset{\sim}{a}'\underset{\sim}{X}) - \mathrm{Var}(\underset{\sim}{a}'\underset{\sim}{X})\mathrm{Var}(Y) = 0$$

with equality if and only if $Y \propto \underset{\sim}{a}'\underset{\sim}{X}$. That is if $Y = (\underset{\sim}{a}'\Sigma\underset{\sim}{a})^{-1/2}\underset{\sim}{a}'\underset{\sim}{X}$. Then

$$B = E(\underset{\sim}{X}\underset{\sim}{X}'\underset{\sim}{a}/(\underset{\sim}{a}'\Sigma\underset{\sim}{a})^{1/2}) = (\underset{\sim}{a}'\Sigma\underset{\sim}{a})^{-1/2}\Sigma\underset{\sim}{a}$$

$$BY = (\underset{\sim}{a}'\Sigma\underset{\sim}{a})^{-1}(\underset{\sim}{a}'\underset{\sim}{X})\Sigma\underset{\sim}{a} \tag{3.34}$$

with this choice of $B\underset{\sim}{Y}$, the residual covariance matrix is

$$\Sigma - BB' = \Sigma - (\underset{\sim}{a}'\Sigma\underset{\sim}{a})^{-1}\Sigma\underset{\sim}{a}\underset{\sim}{a}'\Sigma \tag{3.35}$$

and the loss in (3.33) is zero.

Now let $P\Lambda P'$ be the spectral decomposition of Σ (we assume $|\Sigma| \neq 0$) with P orthogonal and $\Lambda = \text{diag}(\lambda_1,\ldots,\lambda_p)$; $\lambda_1 \geq \cdots \geq \lambda_p$. We may write $\underset{\sim}{a} = cP\Lambda^{-1/2}\underset{\sim}{r}$ where c is a constant and $\underset{\sim}{r} \in R^p$, $\underset{\sim}{r}'\underset{\sim}{r} = 1$. Then, BY may be expressed as

$$BY = (\underset{\sim}{a}'\Sigma\underset{\sim}{a})^{-1}(\underset{\sim}{a}'\underset{\sim}{X})\Sigma\underset{\sim}{a} = P\Lambda^{1/2}\underset{\sim}{r}\underset{\sim}{r}'\Lambda^{-1/2}P'\underset{\sim}{X} = P\Lambda^{1/2}R\begin{bmatrix}1 & 0'\\ 0 & 0\end{bmatrix}R'\Lambda^{-1/2}P'\underset{\sim}{X} \tag{3.36}$$

where $R = (\underset{\sim}{r},\underset{\sim}{r}_2,\ldots,\underset{\sim}{r}_p)$ is orthogonal. Consequently BY is in the class Γ_1 defined in Theorem 3.2. That is, over all possible predictors, the loss (3.33) is minimized when BY is as in (3.34) which is a member of Γ_1.

As a further specialization, suppose our main concern is variable one. To reflect this, we may take $a' = (1,0,\ldots,0)$ so that (3.33) equals entry (1,1) of the residual covariance matrix. In this situation, we may simply take $(X_1,0,\ldots,0)$ as our predictor and have $\underset{\sim}{X} - (X_1,\ldots,0)' = (0,X_2,\ldots,X_p)'$, which has residual covariance matrix with entry (1,1) equal to zero. However the predictor $(X_1,0,\ldots,0)'$ does nothing to predict the remaining variables and conceivably, one can find a predictor that is as good at predicting X_1 and yet gives better prediction of the other variables. Any other predictor will give a residual covariance matrix with entry (1,1) greater than zero, unless the predictor's first entry is (a.s.) X_1. We are then limited to predictors of the form AX_1 with A a $p \times 1$ vector having 1 as the first entry. From (3.34), we obtain the predictor

$$\frac{1}{\sigma_{11}}\Sigma_{(1)}X_1 \tag{3.37}$$

where $\Sigma_{(1)}$ is the first column of Σ, and σ_{11} its (1,1) entry. This predictor also gives a residual covariance matrix with entry (1,1) equal to zero.

We conclude our discussion by showing that, among all predictors AX_1 with first entry X_1, (3.37) gives a residual covariance matrix whose entries are smaller than

or equal to those corresponding to the residual covariance matrix given by any other predictor of the prescribed form.

To this end let a predictor AX_1 be written as

$$\underset{\sim}{X} = \begin{bmatrix} X_1 \\ \underset{\sim}{Z} \end{bmatrix}, \quad AX_1 = \begin{bmatrix} X_1 \\ CX_1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_{11} & \underset{\sim}{\sigma}_1' \\ \underset{\sim}{\sigma}_1 & \Sigma_{11} \end{bmatrix}$$

then $\underset{\sim}{X} - AX_1$ has residual covariance matrix

$$\begin{bmatrix} 0 & 0' \\ 0 & \Sigma_{11} - \underset{\sim}{\sigma}_1 C' - C\underset{\sim}{\sigma}_1' + C\sigma_{11}C' \end{bmatrix}$$

but

$$\Sigma_{11} - \underset{\sim}{\sigma}_1 C' - C\underset{\sim}{\sigma}_1' + C\sigma_{11}C' = \left[\frac{\underset{\sim}{\sigma}_1}{\sqrt{\sigma_{11}}} - \sqrt{\sigma_{11}}C\right]\left[\frac{\underset{\sim}{\sigma}_1}{\sqrt{\sigma_{11}}} - \sqrt{\sigma_{11}}C\right]' + \Sigma_{11} - \frac{1}{\sigma_{11}}\underset{\sim}{\sigma}_1\underset{\sim}{\sigma}_1'$$

$$\geq \Sigma_{11} - \frac{1}{\sigma_{11}}\underset{\sim}{\sigma}_1\underset{\sim}{\sigma}_1' \ .$$

4. Introducing the Principal Components Sequentially

Here we extend Theorem 3.1 to the sequential selection of approximators. Suppose $k = 1$, so that the covariance matrices are of the form $E(\underset{\sim}{X} - AY)(\underset{\sim}{X} - AY)'$ where A is a $p \times 1$ vector of constants and Y a univariate random variable. In this section, we sometimes write $\lambda[M]$ to denote that λ is an eigenvalue of M.

Let $\underset{\sim}{P}_1$ be an eigenvector of Σ, corresponding to its largest eigenvalue, and $A_1Y_1 = \underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X}$. Theorem 3.1 establishes that this AY has the property that the ith largest eigenvalue of the corresponding residual covariance matrix $E(\underset{\sim}{X} - \underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X})(\underset{\sim}{X} - \underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X})'$ is less than or equal to the corresponding eigenvalue of any possible residual covariance matrix $(i = 1,\ldots,p)$.

Proceeding in a sequential manner, we seek A_2Y_2 uncorrelated with A_1Y_1 with the same minimizing property among the class of AY's uncorrelated with A_1Y_1. Then A_3Y_3 uncorrelated with A_1Y_1 and A_2Y_2, and with the minimizing property is desired, and so on. The next theorem tells us that principal components are the solution.

Theorem 4.1. Let $\underset{\sim}{X}$ be a $p \times 1$ random vector with zero expectation and covariance matrix Σ of rank r. Also let $P = \{AY | A$ is a $p \times 1$ vector and Y a random variable with $EY = 0$; $EY^2 = 1\}$. For $j = 1,\ldots,r$, let $A_jY_j \in P$. Then,

$$\lambda_i[E(\underset{\sim}{X} - A_jY_j)(\underset{\sim}{X} - A_jY_j)'] \leq \lambda_i[E(\underset{\sim}{X} - AY)(\underset{\sim}{X} - AY)'] \quad (i = 1,\ldots,p) ,$$

for every AY in P uncorrelated with $A_1Y_1,\ldots,A_{j-1}Y_{j-1}$, if and only if $A_jY_j = \underset{\sim}{P}_j\underset{\sim}{P}_j'\underset{\sim}{X}$ where $\underset{\sim}{P}_j$ is an eigenvector of Σ corresponding to the eigenvalue $\lambda_j[\Sigma]$.

Proof. For $j = 1$, application of Theorem 3.1 provides $A_1Y_1 = \underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X}$ where $\underset{\sim}{P}_1$ is as stated. For $j = 2$. Let $P = (\underset{\sim}{P}_1 : C) = (\underset{\sim}{P}_1, \underset{\sim}{P}_2, \ldots, \underset{\sim}{P}_p)$ be orthogonal where $\underset{\sim}{P}_2,\ldots,\underset{\sim}{P}_p$ are eigenvectors of Σ corresponding respectively to $\lambda_2[\Sigma],\ldots,\lambda_p[\Sigma]$, so that $\underset{\sim}{X} = PP'\underset{\sim}{X} = \underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X} + CC'\underset{\sim}{X}$ and

$$\begin{aligned} E(\underset{\sim}{X} - A_2Y_2)(\underset{\sim}{X} - A_2Y_2)' &= E(\underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X} + CC'\underset{\sim}{X} - A_2Y_2)(\underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X} + CC'\underset{\sim}{X} - A_2Y_2)' \\ &= E(\underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X})(\underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X})' + E(CC'\underset{\sim}{X} - A_2Y_2)(CC'\underset{\sim}{X} - A_2Y_2)' \\ &= \lambda_1\underset{\sim}{P}_1\underset{\sim}{P}_1' + E(CC'\underset{\sim}{X} - A_2Y_2)(CC'\underset{\sim}{X} - A_2Y_2)' \end{aligned}$$

where the cross terms are zero because $\underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X}$ is uncorrelated with $CC'\underset{\sim}{X}$ and A_2Y_2. The result now follows from Theorem 3.1 by noting that $CC'\underset{\sim}{X}$ has covariance matrix

$$CC'\Sigma CC' = CC'(\underset{\sim}{P}_1 : C)\Lambda\begin{bmatrix}\underset{\sim}{P}_1' \\ C'\end{bmatrix}CC' = \sum_{2}^{p}\lambda_i\underset{\sim}{P}_i\underset{\sim}{P}_i' = C\begin{pmatrix}\lambda_2 & & \\ & \ddots & \\ & & \lambda_p\end{pmatrix}C' .$$

Alternatively, if $B_2 = E(CC'\underset{\sim}{X}Y_2) = CC'E(\underset{\sim}{X}Y_2) = CC'\underset{\sim}{u}$, say. From Lemma 2.1, we have

$$\begin{aligned} \lambda_1\underset{\sim}{P}_1\underset{\sim}{P}_1' + E(CC'\underset{\sim}{X} - A_2Y_2)(CC'\underset{\sim}{X} - A_2Y_2)' &\geq \lambda_1\underset{\sim}{P}_1\underset{\sim}{P}_1' + CC'\Sigma CC' - B_2B_2' \\ &= \lambda_1\underset{\sim}{P}_1\underset{\sim}{P}_1' + CC'(\Sigma - \underset{\sim}{u}\underset{\sim}{u}')CC' . \end{aligned} \tag{4.1}$$

Since $C'\underset{\sim}{P}_1 = 0$, by simply multiplying (4.1) on the right by $\underset{\sim}{P}_1$ we see that $\underset{\sim}{P}_1$ is an eigenvector corresponding to the eigenvalue λ_1. Similarly any eigenvector of the second term, perpendicular to $\underset{\sim}{P}_1$, is also an eigenvector of the whole right hand side of (4.1) for the same eigenvalue. Thus to minimize the eigenvalues of (4.1) we need only minimize the eigenvalues of the second term.

But

$$CC'(\Sigma - \underset{\sim}{u}\underset{\sim}{u}')CC' = E(CC'\underset{\sim}{X} - A_2Y_2)(CC'\underset{\sim}{X} - A_2Y_2)'$$

and $CC'\underset{\sim}{X}$ is a vector with zero expectation and covariance matrix $CC'\Sigma CC'$. From the result for $j = 1$, we must take $A_2Y_2 = \underset{\sim}{u}_2\underset{\sim}{u}_2'\underset{\sim}{X}$ where $\underset{\sim}{u}_2$ is an eigenvector of $CC'\Sigma CC'$. Thus $A_2Y_2 = \underset{\sim}{P}_2\underset{\sim}{P}_2'\underset{\sim}{X}$ gives the result for $j = 2$. Similarly for $j = 2,\ldots,r$. □

Remark. The procedure was only carried through r stages where r = rank of Σ, instead of p stages. The reason for this is that $\underset{\sim}{X} = \underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X} + \cdots + \underset{\sim}{P}_r\underset{\sim}{P}_r'\underset{\sim}{X}$ (a.s.) so that any variable uncorrelated with $\underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X},\ldots,\underset{\sim}{P}_r\underset{\sim}{P}_r'\underset{\sim}{X}$ is uncorrelated with $\underset{\sim}{X}$ and then

$$E(\underset{\sim}{X} - A\underset{\sim}{Y})(\underset{\sim}{X} - A\underset{\sim}{Y})' = \Sigma + AE(Y^2)A' \geq \Sigma .$$

If one insists on introducing p components, then since (a.s.) $\underset{\sim}{P}_{r+1}'\underset{\sim}{X} = \cdots = \underset{\sim}{P}_p'\underset{\sim}{X} = 0$, $\underset{\sim}{P}_{r+1}\underset{\sim}{P}_{r+1}'\underset{\sim}{X},\ldots,\underset{\sim}{P}_p\underset{\sim}{P}_p'\underset{\sim}{X}$ may be taken as the extra components. However no matter how the extra components are introduced, stages $r + 1,\ldots,p$ are irrelevant.

Remark. An alternative way of sequentially introducing the first r (= rank of Σ) principal components is to begin as above with $A_1Y_1 = \underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X}$ having the property that its eigenvalues are less than or equal to the corresponding eigenvalues of any possible residual covariance matrix. However, as a second step, we consider the residual $\underset{\sim}{X} - \underset{\sim}{P}_1\underset{\sim}{P}_1'\underset{\sim}{X} = \underset{\sim}{Z}$ which itself has zero expectation and covariance matrix $\Sigma - \lambda_1\underset{\sim}{P}_1\underset{\sim}{P}_1' = \sum_{i=2}^{p} \lambda_i\underset{\sim}{P}_i\underset{\sim}{P}_i'$. *Applying Theorem 3.1 to the residual* $\underset{\sim}{Z}$, *one finds that* $AY = \underset{\sim}{P}_2\underset{\sim}{P}_2'\underset{\sim}{X}$ gives a corresponding residual covariance matrix

$$E(\underset{\sim}{Z} - \underset{\sim}{P}_2\underset{\sim}{P}_2'\underset{\sim}{X})(\underset{\sim}{Z} - \underset{\sim}{P}_2\underset{\sim}{P}_2'\underset{\sim}{X})' = \Sigma - \lambda_1\underset{\sim}{P}_1\underset{\sim}{P}_1' - \lambda_2\underset{\sim}{P}_2\underset{\sim}{P}_2'$$

such that for $i = 1,\ldots,p$; its ith largest eigenvalue is less or equal than the corresponding eigenvalue of any other possible residual covariance matrix. In this way, the first r principal components may be introduced sequentially in terms of approximating successive residuals. We then have

Corollary 4.2. Suppose we select the approximators one at a time from the residual covariance of the previous stage. Let $\underset{\sim}{Z}_q = \underset{\sim}{X} - \sum_{i=1}^{q} \underset{\sim}{P}_i\underset{\sim}{P}_i'\underset{\sim}{X}$ be the residual matrix after q steps. Then the choice $AY = \underset{\sim}{P}_{q+1}\underset{\sim}{P}_{q+1}'\underset{\sim}{X}$, for stage $q + 1$, minimizes the eigenvalues of $E(\underset{\sim}{Z}_q - A\underset{\sim}{Y})(\underset{\sim}{Z}_q - A\underset{\sim}{Y})'$.

5. Some Sample Interpretations

5.1. p-Space Interpretations

Let $(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n) = \chi$ denote n observations on the $p \times 1$ random vector $\underset{\sim}{X}$. We can think of the n observations as n points in R^p. Moreover, one can assign probability $1/n$ to each observation vector and apply the results from the population model to obtain the sample results. We intend, however, to go in the reverse direction and show how Pearson's (1901) result may be extended to hold for all increasing functions of eigenvalues. This frames the result by Okamoto and Kanazawa so that it may be viewed as a natural population analog. In our development, although we speak of projections on planes not passing through the origin we really make suitable translations and take proper projections on subspaces.

Suppose that we are interested in finding that line which best fits the n given points, in the sense that the sum of the squares of the distances from these n points to the line is a minimum. More generally, we ask for the hyperplane of dimension $k(\leq p)$ which best fits the data in the above sum of squares sense.

Let $\bar{\underset{\sim}{x}} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{x}_i$, be the centroid of the n points and $S = \sum_{i=1}^{n} (\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})'$ the cross products matrix. We have the following theorem due to Pearson (1901).

Theorem 5.1 [Pearson]. The hyperplane of dimension $k(\leq p)$ such that the sum of squares of the distances from the points to the plane is a minimum is of the form

$$\{\underset{\sim}{x} | \underset{\sim}{x} = \bar{\underset{\sim}{x}} + P\underset{\sim}{\mu}, \underset{\sim}{\mu} \in R^k\}$$

where $P = (\underset{\sim}{P}_1,\ldots,\underset{\sim}{P}_k)$ and its k columns constitute a set of k orthonormal eigenvectors of the cross products matrix S corresponding respectively to the k largest eigenvalues. □

If $\bar{\underset{\sim}{x}} + P\underset{\sim}{\mu}_i$ is the point in the plane closest to $\underset{\sim}{x}_i$ then

$$P\underset{\sim}{\mu}_i = PP'(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}}) = \underset{\sim}{P}_1\underset{\sim}{P}_1'(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}}) + \cdots + \underset{\sim}{P}_k\underset{\sim}{P}_k'(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}}) \tag{5.1}$$

and if $\lambda_1 \geq \cdots \geq \lambda_p$ are the eigenvalues of S then the sum of squared distances from the points $\underset{\sim}{x}_i$ to the hyperplane equals

$$SS(\text{distances}) = \lambda_{k+1} + \cdots + \lambda_p . \tag{5.2}$$

Pearson's result then tells us that the "best" hyperplane is determined by $\bar{\underset{\sim}{x}}, \underset{\sim}{P}_1, \ldots, \underset{\sim}{P}_k$ and the point nearest to $\underset{\sim}{x}_i$ is $\bar{\underset{\sim}{x}} + P_1 P_1'(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}}) + \cdots + P_k P_k'(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})$.

We may also think of $\underset{\sim}{P}_i$ as determining a privileged direction in R^p. The line $\{\underset{\sim}{P}_i \alpha \mid \alpha \in R\}$ is the line perpendicular to $\underset{\sim}{P}_1, \ldots, \underset{\sim}{P}_{i-1}$ which best fits the points $(\underset{\sim}{x}_1 - \bar{\underset{\sim}{x}}, \ldots, \underset{\sim}{x}_n - \bar{\underset{\sim}{x}}) = \chi^*$ or equivalently the line that best fits (no restrictions now) the residuals $\chi^* - \underset{\sim}{P}_1 \underset{\sim}{P}_1' \chi^* \cdots - \underset{\sim}{P}_{i-1} \underset{\sim}{P}_{i-1}' \chi^*$. We now show how Pearson's result can be extended to a statement about eigenvalues.

Let P be a $p \times k$ matrix, $\mu = (\underset{\sim}{\mu}_1, \ldots, \underset{\sim}{\mu}_n)$ be a $k \times n$ matrix and $\underset{\sim}{x}$ a $p \times 1$ vector, then the points $\underset{\sim}{x} + \underset{\sim}{P}\underset{\sim}{\mu}_i$ $(i = 1, \ldots, n)$ all lie in the hyperplane parallel to that spanned by the columns of P. Consider the matrix

$$\sum_{i=1}^{n} (\underset{\sim}{x}_i - \underset{\sim}{x} - \underset{\sim}{P}\underset{\sim}{\mu}_i)(\underset{\sim}{x}_i - \underset{\sim}{x} - \underset{\sim}{P}\underset{\sim}{\mu}_i)' \,. \tag{5.3}$$

If $\underset{\sim}{x} + \underset{\sim}{P}\underset{\sim}{\mu}_i$ is the projection of $\underset{\sim}{x}_i$ over the hyperplane then the trace of (5.3) gives us the sum of squared distances from the $\underset{\sim}{x}_i$'s to the hyperplane and Pearson's result tells us how to select $\underset{\sim}{x}$ and $\underset{\sim}{P}$. Here we show how to make the eigenvalues of (5.3) as small as possible by a correct choice of $\underset{\sim}{x}, P, \underset{\sim}{\mu}_1, \ldots, \underset{\sim}{\mu}_n$. It is clear that the columns of P can be taken to be orthonormal in the following optimization.

Theorem 5.2. To minimize the eigenvalues of the matrix

$$\sum_{i=1}^{n} (\underset{\sim}{x}_i - \underset{\sim}{x} - \underset{\sim}{P}\underset{\sim}{\mu}_i)(\underset{\sim}{x}_i - \underset{\sim}{x} - \underset{\sim}{P}\underset{\sim}{\mu}_i)' \tag{5.4}$$

$\underset{\sim}{x}$ may be selected as $\bar{\underset{\sim}{x}}$, and

$$P(\underset{\sim}{\mu}_1, \ldots, \underset{\sim}{\mu}_n) = PP'\chi^*$$

where χ^* stands for the matrix χ with each of its rows corrected for the mean and $\underset{\sim}{P}_1, \ldots, \underset{\sim}{P}_k$ constitute a set of k orthonormal eigenvectors of the cross products matrix S corresponding respectively to the k largest eigenvalues $\lambda_1 \geq \cdots \geq \lambda_k$. Thus, any increasing function of the eigenvalues is minimized by this choice.

Proof. Without loss of generality, we may assume that $P\bar{\mu} = \sum_{i=1}^{n} \underset{\sim}{P}\underset{\sim}{\mu}_i = \underset{\sim}{0}$ for, if this is not the case, $\underset{\sim}{x}$ may be replaced by $\underset{\sim}{x} + P\bar{\underset{\sim}{\mu}}$ and $\underset{\sim}{P}\underset{\sim}{\mu}_i$ by $\underset{\sim}{P}\underset{\sim}{\mu}_i - P\bar{\underset{\sim}{\mu}}$. Suppose then, $P\bar{\mu} = 0$. We have the non-negative definite ordering

$$\sum_i (\underset{\sim}{x}_i - \underset{\sim}{x} - P\underset{\sim}{\mu}_i)(\underset{\sim}{x}_i - \underset{\sim}{x} - P\underset{\sim}{\mu}_i)' = \sum_i (\underset{\sim}{x}_i - \bar{\underset{\sim}{x}} - P\underset{\sim}{\mu}_i + \bar{\underset{\sim}{x}} - \underset{\sim}{x})(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}} - P\underset{\sim}{\mu}_i + \bar{\underset{\sim}{x}} - \underset{\sim}{x})'$$

$$= \sum_i (\underset{\sim}{x}_i - \bar{\underset{\sim}{x}} - P\underset{\sim}{\mu}_i)(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}} - P\underset{\sim}{\mu}_i)' + \Sigma(\bar{\underset{\sim}{x}} - \underset{\sim}{x})(\bar{\underset{\sim}{x}} - \underset{\sim}{x})'$$

$$\geq \sum_i (\underset{\sim}{x}_i - \bar{\underset{\sim}{x}} - P\underset{\sim}{\mu}_i)(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}} - P\underset{\sim}{\mu}_i)' \qquad (5.5)$$

with strict inequality unless $\underset{\sim}{x} = \bar{\underset{\sim}{x}}$ so that $\underset{\sim}{x}$ must be taken as the centroid of $\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n$.

Next we write $Y = (\underset{\sim}{y}_1,\ldots,\underset{\sim}{y}_n) = (\underset{\sim}{x}_1 - \bar{\underset{\sim}{x}},\ldots,\underset{\sim}{x}_n - \bar{\underset{\sim}{x}})$ and obtain

$$\sum_{i=1}^{n} (\underset{\sim}{y}_i - P\underset{\sim}{\mu}_i)(\underset{\sim}{y}_i - P\underset{\sim}{\mu}_i)' = (Y - P\underset{\sim}{\mu})(Y - P\underset{\sim}{\mu})', \quad \underset{\sim}{\mu} = (\underset{\sim}{\mu}_1 \vdots \cdots \vdots \underset{\sim}{\mu}_r) .$$

The above matrix has the same nonzero eigenvalues as the matrix

$$(Y - P\underset{\sim}{\mu})'(Y - P\underset{\sim}{\mu})$$

and we have

$$(Y - P\underset{\sim}{\mu})'(Y - P\underset{\sim}{\mu}) = (Y - PP'Y + PP'Y - P\underset{\sim}{\mu})'(Y - PP'Y + PP'Y - P\underset{\sim}{\mu})$$

$$= (Y - PP'Y)'(Y - PP'Y) + (PP'Y - P\underset{\sim}{\mu})'(PP'Y - P\underset{\sim}{\mu})$$

$$\geq (Y - PP'Y)'(Y - PP'Y)$$

with strict inequality unless $PP'Y = P\underset{\sim}{\mu}$ or $P'Y = \underset{\sim}{\mu}$. With this choice we have

$$(Y - PP'Y)'(Y - PP'Y) = Y'(I - PP')Y .$$

We want to minimize the eigenvalues of this matrix or equivalently of the matrix

$$(I - PP')YY'(I - PP') = (I - PP')S(I - PP') = QQ'SQQ'$$

where Q is such that $(P \vdots Q)$ is orthogonal. Next, consider the product

$$\begin{bmatrix} P' \\ Q' \end{bmatrix} QQ'SQQ'(P \quad Q) = \begin{bmatrix} 0 & 0 \\ 0 & Q'SQ \end{bmatrix} \qquad (5.7)$$

which has the same eigenvalues as $QQ'SQQ'$. Its eigenvalues are k zeroes and those of the lower right corner of

$$\begin{bmatrix} P' \\ Q' \end{bmatrix} S(P \quad Q) = \begin{bmatrix} P'SP & P'SQ \\ Q'SP & Q'SQ \end{bmatrix} .$$

The Poincare separation theorem (c.f. Bellman (1970), p. 117) shows that

$$\lambda_1[Q'SQ] \geq \lambda_{k+1}[S],\ldots,\lambda_{p-k}[Q'SQ] \geq \lambda_p[S]$$

with at least one strict inequality unless an orthogonal matrix of the form

$$\begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix}$$

with R_1 of order k exists, such that

$$\begin{bmatrix} R_1' & 0 \\ 0 & R_2' \end{bmatrix} \begin{bmatrix} P' \\ Q' \end{bmatrix} S (P \quad Q) \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix} = \text{diag}(\lambda_1, \ldots, \lambda_p)$$

where $\lambda_1 \geq \cdots \geq \lambda_p$ are the eigenvalues of S.

Then the columns of PR_1 must constitute a set of k orthonormal eigenvectors of S corresponding respectively to the k largest eigenvalues $\lambda_1 \geq \cdots \geq \lambda_k$. Finally $PP' = PR_1R_1'P'$ gives the desired result. □

We now state the sample version of the result by Okamoto and Kanazawa (1968) or Okamoto (1969). Our intention is to illustrate how this is connected to Pearson (1901).

Theorem 5.3. Let $\tilde{\underset{\sim}{X}}$ denote the random variable which assumes each of the observed sample values $\underset{\sim}{x}_i$, $i = 1,2,\ldots,n$ with probability $\frac{1}{n}$ and $\tilde{\underset{\sim}{U}}$ denote any corresponding $k \times 1$ vector. Then, the eigenvalues of

$$E[\tilde{\underset{\sim}{X}} - \bar{\underset{\sim}{x}} - A\tilde{\underset{\sim}{U}}][\tilde{\underset{\sim}{X}} - \bar{\underset{\sim}{x}} - A\tilde{\underset{\sim}{U}}]' = \frac{1}{n} \sum_{i=1}^{n} (\underset{\sim}{x}_i - \bar{\underset{\sim}{x}} - A\underset{\sim}{u}_i)(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}} - A\underset{\sim}{u}_i)'$$

are simultaneously minimized over all A of dimension $p \times k$ and all sets of values $\{\underset{\sim}{u}_i\}$ for $\tilde{\underset{\sim}{U}}$ by

$$A\underset{\sim}{u}_i = (\underset{\sim}{P}_1\underset{\sim}{P}_1' + \cdots + \underset{\sim}{P}_k\underset{\sim}{P}_k')(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})$$

where $\underset{\sim}{P}_1, \ldots, \underset{\sim}{P}_k$ is a set of orthonormal eigenvectors of the matrix $\frac{1}{n} \sum_{i=1}^{n} (\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})'$ corresponding to the k largest eigenvectors.

Proof. The result follows directly from Theorem 5.2 or as a special case of Theorem 3.1. □

We also have another geometric interpretation in terms of approximating cross product matrices.

Let a hyperplane be $\{\underset{\sim}{v} | \underset{\sim}{v} = \bar{\underset{\sim}{x}} + P\underset{\sim}{\mu};\ \underset{\sim}{\mu} \in R^k\}$ where $P = (\underset{\sim}{P}_1, \ldots, \underset{\sim}{P}_k)$ is $p \times k$ with orthonormal columns. The projection of a given point $\underset{\sim}{x}_i$ into this hyperplane is the point $\underset{\sim}{v}_i = \bar{\underset{\sim}{x}} + PP'(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})$. Let $S = \sum_{i=1}^{n} (\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})'$ denote the cross products matrix and let

$$S_v = \sum_{1}^{n} (\underset{\sim}{v}_i - \bar{\underset{\sim}{v}})(\underset{\sim}{v}_i - \bar{\underset{\sim}{v}})' = \sum_{1}^{n} PP'(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})'PP' = PP'SPP' \tag{5.8}$$

denote the cross products matrix of the projected points. We want to find a hyperplane in R^p, passing through the centroid $\bar{x}$, and such that S_v is close to S.

Theorem 5.4. In order to minimize the eigenvalues of the discrepancy in sums of cross product matrices

$$S - S_v = S - PP'SPP' ,$$

over all k dimensional planes passing through $\bar{x}$, we must take the plane where the columns of P are a set of eigenvectors of S corresponding to the k largest eigenvalues.

Proof. Let Q be such that $(P:Q)$ is orthogonal. The eigenvalues of $S - PP'SPP'$ are exactly the same as those of

$$\begin{bmatrix} P' \\ Q' \end{bmatrix} S(P \quad Q) - \begin{bmatrix} P' \\ Q' \end{bmatrix} PP'SPP'(P \quad Q) = \begin{bmatrix} 0 & P'SQ \\ Q'SP & Q'SQ \end{bmatrix}$$

whose eigenvalues are k zeroes together with those of $Q'SQ$.

From the argument given after equation (5.7), it follows that the hyperplane has to be of the form

$$\{v \mid v = \bar{x} + Pu;\ u \in R^k\} \tag{5.9}$$

where $P = (P_1,\ldots,P_k)$ and $P_1,\ldots,P_k$ constitute a set of orthonormal eigenvectors of the matrix S corresponding respectively to the k largest eigenvalues $\lambda_1 \geq \cdots \geq \lambda_k$. □

Remark. If, alternatively, we consider the residuals $x_i - v_i$ with v_i as in (5.8), then the cross products matrix for the residuals is

$$S_r = \Sigma(x_i - v_i)(x_i - v_i)' = \Sigma(x_i - \bar{x} - PP'(x_i - \bar{x}))(x_i - \bar{x} - PP'(x_i - \bar{x}))'$$

and Theorem 5.2 tells us that, to minimize the eigenvalues of S_r, the hyperplane must also be taken as in (5.9).

5.5. n-Space Interpretation

A geometrical interpretation of principal components in n space does not seem to have been given by previous workers. Let $\chi = (x_1,\ldots,x_n)$ denote n observations on the p-vector X. If the i-th row of $\chi = (x_1,\ldots,x_n)$ is denoted by y_i' and the i-th row of $\chi^* = (x_1 - \bar{x},\ldots,x_n - \bar{x})$ by $y_i^{*\prime}$, then these rows are points in R^n.

With the mean corrected rows of the data matrix, it is natural to apply Euclidean distances and inner products. Suppose then, we ask for the plane of dimension $k(\leq p)$ which best fits the p points determined by the rows of χ^* and which passes through the origin. If

$$C = (\underset{\sim}{c}_1,\ldots,\underset{\sim}{c}_n) = \begin{bmatrix} \underset{\sim}{r}_1' \\ \vdots \\ \underset{\sim}{r}_p' \end{bmatrix}$$

denotes the points in the hyperplane with $\underset{\sim}{r}_i'$ being the point closest to $\underset{\sim}{y}_i^{*\prime}$, we want to minimize the sum of the squared norms of the rows of the matrix

$$\begin{bmatrix} \underset{\sim}{y}_1^{*\prime} - \underset{\sim}{r}_1' \\ \vdots \\ \underset{\sim}{y}_p^{*\prime} - \underset{\sim}{r}_p' \end{bmatrix} = (\chi^* - C) = (\underset{\sim}{x}_1^* - \underset{\sim}{c}_1,\ldots,\underset{\sim}{x}_n^* - \underset{\sim}{c}_n) \tag{5.10}$$

which is clearly equivalent to minimizing the sum of squares of all the entries of the above matrix or to minimizing the sum of the squared norms of the columns. From Pearson's result, the best choice of C is then

$$C = \underset{\sim}{P}_1\underset{\sim}{P}_1'\chi^* + \cdots + \underset{\sim}{P}_k\underset{\sim}{P}_k'\chi^* \tag{5.11}$$

where $\underset{\sim}{P}_1,\ldots,\underset{\sim}{P}_k$ are as in (5.1).

For the case $k = 1$, we can consider the problem as one of first projecting the rows of χ^* on any $n \times 1$ vector $\underset{\sim}{a}'$. These projections are given by the rows of

$$\chi^*\underset{\sim}{a}\underset{\sim}{a}'/(\underset{\sim}{a}'\underset{\sim}{a}) = C_{\underset{\sim}{a}} \tag{5.12}$$

where $C_{\underset{\sim}{a}}$ is of rank 1 when rank $(\chi^*) \geq 1$. Thus, the sum of squares of the i-th row of $\chi^* - C_{\underset{\sim}{a}}$ is the squared distance from $\underset{\sim}{y}_i^*$ to the line determined to $\underset{\sim}{a}$. Moreover, we know that the choice $\underset{\sim}{a}' = \underset{\sim}{P}_1'\chi^* = (\underset{\sim}{P}_1'(\underset{\sim}{x}_1 - \bar{\underset{\sim}{x}}),\ldots,\underset{\sim}{P}_1'(\underset{\sim}{x}_n - \bar{\underset{\sim}{x}}))$, the sample values of the first principal component, produces $C_{\underset{\sim}{a}}$ of the form (5.12) or

$$\chi^*\chi^*{}''\underset{\sim}{P}_1\underset{\sim}{P}_1'\chi^*/\underset{\sim}{P}_1'\chi^*\chi^*\underset{\sim}{P}_1 = S\underset{\sim}{P}_1\underset{\sim}{P}_1'\chi^*/\underset{\sim}{P}_1'S\underset{\sim}{P}_1 = \lambda_1\underset{\sim}{P}_1\underset{\sim}{P}_1'\chi^*/\lambda_1 = \underset{\sim}{P}_1\underset{\sim}{P}_1'\chi^*$$

conforming to the optimal choice of C given by (5.11) with $k = 1$.

Remark. The first principal component minimizes the sum of squared distances to the rows of the mean corrected data matrix χ^*. This is illustrated in Figure 5.1. Our geometrical interpretation helps us understand the relative importance of a row with large length (sample variance) in determining the principal component.

Remark. If the rows of χ^* are first scaled to become unit vectors, then extracting principal components from the new matrix is equivalent to using the sample correlation matrix R. As the vectors in the geometric interpretation are now all of unit length, minimizing the sum of squared errors is equivalent to maximizing (minimizing) the sum of $\cos^2\theta_i$ ($\sin^2\theta_i$) where θ_i is the angle between $\underset{\sim}{y}_i^*$ and $\underset{\sim}{a}$.

It is easy to see that $\underset{\sim}{P}_i'\chi^*$ determines a privileged line in n space. It is the line perpendicular to $\underset{\sim}{P}_1'\chi^*,\ldots,\underset{\sim}{P}_{i-1}'\chi^*$ that passes through the origin and which best fits the rows of χ^* or equivalently the line through the origin which best fits the rows (no restrictions now) of the residual $\chi^* - \underset{\sim}{P}_1\underset{\sim}{P}_1'\chi^* \cdots -\underset{\sim}{P}_{i-1}\underset{\sim}{P}_{i-1}'\chi^*$.

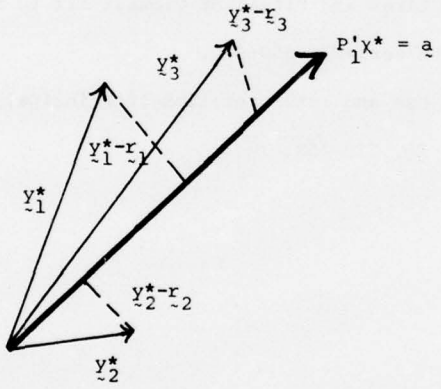


Figure 5.1. Case p = 3, showing sample first principal component minimizing the sum of squared distances

$$\|\underset{\sim}{y}_1^* - \underset{\sim}{r}_1\|^2 + \|\underset{\sim}{y}_2^* - \underset{\sim}{r}_2\|^2 + \|\underset{\sim}{y}_3^* - \underset{\sim}{r}_3\|^2$$

from the $\underset{\sim}{y}_i^*$'s to line.

Acknowledgment

The authors wish to thank Dr. M. El-Bassiouni for suggesting improvements in the presentation of this material.

REFERENCES

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. New York, Wiley.

Bellman, R. (1970). An Introduction to Matrix Analysis. McGraw-Hill, New York.

Darroch, J. N. (1965). An Optimal Property of Principal Components. Ann. Math. Statist., 36, 1579-1582.

Girshick, M. A. (1936). Principal Components. J. Amer. Statist. Assoc., 31, 519-528.

Hotelling, H. (1933). Analysis of Complex Statistical Variables into Principal Components. J. Educ. Psych., 24, 417-444, 498-520.

Okamoto, M. (1969). Optimality of Principal Components, Multivariate Analysis II, 673-685, Ed. P. R. Krishnaiah, Academic Press, New York.

Okamoto, M. and Kanazawa M. (1968). Minimization of Eigenvalues of a Matrix and Optimality of Principal Components. Ann. Math. Statist., 39, 859-863.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Phil. Mag., 2, (sixth series)*, 559-572.

Rao, C. R. (1964). The Use and Interpretation of Principal Component Analysis in Applied Research. Sankhya, 26, 329-358.

RH/RAJ/scr

(14) MRC-TSR-1836

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 1836 (9) Technical summary rept.
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): (6) SOME OPTIMAL PROPERTIES AND INTERPRETATIONS OF PRINCIPAL COMPONENTS.

5. TYPE OF REPORT & PERIOD COVERED: Summary Report - no specific reporting period

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): (10) Raul Hudlet [redacted] Richard A. Johnson

8. CONTRACT OR GRANT NUMBER(s): (15) DAAG29-75-C-0024, VNSF-MCS77-09574

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Work Unit # 4 - Probability, Statistics, and Combinatorics

11. CONTROLLING OFFICE NAME AND ADDRESS: See Item 18 below.

12. REPORT DATE: (11) Mar[redacted]978

13. NUMBER OF PAGES: 24

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report): UNCLASSIFIED (12) 12 p.

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, D. C. 20550

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Principal components

Statistical approximations

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Optimal properties are derived and some new geometrical interpretations given for principal components. Typically, our main results concern the simultaneous minimization of eigenvalues of certain covariance matrices which measure the goodness of an approximation. Many popular criteria like total variance and generalized variance, which are increasing functions of the eigenvalues, are then minimized by the best approximator.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED → next page

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

221 200

20. ABSTRACT - Cont'd.

In other situations, the criterion may not be a monotone function of the eigenvalues. ~~In Theorem 3.2, we derive~~ a general optimal class is derived based on the non-negative definite ordering of covariance matrices. ~~Theorem 4.1 gives~~ Also a result is given for the sequential selection of principal components. In the final section, ~~we~~ gives a new geometrical interpretation of the sample principal components.